DOCUMENT RESUME

ED 178 612 TM 009 906

AUTHOR Johnson, Richard T.; Thomas, Wayne P.
TITLE User Experiences in Implementing the RMC Title I Evaluation Models.
PUB DATE Apr 79
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, CA, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Administrative Problems; Compensatory Education Programs; Data Processing; Elementary Secondary Education; *Evaluation Needs; *Guidelines; Models; *Program Administration; Program Development; Program Effectiveness; *Program Evaluation; Student Testing; Testing Problems; *Testing Programs
IDENTIFIERS Elementary Secondary Education Act Title I; *Evaluation Problems; *RMC Models

ABSTRACT

The experiences of state and local education agencies in implementing the RMC evaluation models in their evaluations of Elementary and Secondary Education Act Title I programs are discussed with emphasis on the problems encountered, suggestions for resolving these problems, and encouraging results which have been reported. Four activities are described as common for all of the agencies which have been using the RMC models to measure achievement gains: test selection and administration; scoring and conversion of scores; data analysis; and data aggregation. Problems which have been encountered in implementing the RMC models are classified as (1) procedural--selection of student samples, tests, and RMC model; and communication of results; (2) clerical--conversion of raw scores, comparison of pretest and post test scores, and failure to record testing date; and (3) analytical--errors in data analysis. A total of 27 guidelines are suggested which should be helpful in reducing these problems. Several positive comments regarding the usefulness and potential of the RMC models are included. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178612

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

USER EXPERIENCES IN IMPLEMENTING THE RMC TITLE I EVALUATION MODELS

Richard T. Johnson and Wayne P. Thomas

Virginia Polytechnic Institute and State University

April 1979

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R.T. Johnson

Wayne P. Thomas

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the

Annual Meeting of the American Educational Research Association

San Francisco, California, April 7-13, 1979

Session 24.24

IM009 906

Many school systems in the country have made progress in testing and implementing the ESEA Title I Evaluation and Reporting System. During the 1976-77 school year, 28 of the SEA's across the nation had LEA's using one or more of the proposed models to evaluate the reading and/or math outcomes attributable to participation in a Title I project. Most of these LEA's were using the techniques for the first time. Information gleaned from the evaluation personnel in these LEA's gives an idea of the administrative procedures involved in effecting the change to new methods. The Title I staffs in the states with the longest and most comprehensive experience in using the new procedures (Florida, Iowa, Maine, Ohio, South Carolina, and South Dakota) have shared the results of their experiences in implementing the Title I evaluation models and have done so in particularly helpful ways. State Title I officials and the local staffs of more than 20 school divisions in the Commonwealth of Virginia have also been very cooperative in sharing their experiences in implementing the models.

The legal mandate for this "field test" and documentation effort is in Subsection F of Section 151 of ESEA, Title I, which states that the Commissioner of Education must require SEA's and LEA's to use "techniques... and methodology...for producing data which are comparable on a statewide and nationwide basis." The development and publication of valid evaluation models as required by Section 151 is the first step toward the nationwide production of comparable data; subsequent necessary steps are the monitoring of the use of the models and the development of means to make their use as error-free as possible.

The purpose of this paper is to describe the problems experienced by state and local personnel who used the evaluation models to assess project effects, to suggest ways in which these problems can be solved, and to mention some very encouraging results that have come about through the use of the models.

There are many ways in which LEA's and SEA's have gathered, analyzed, aggregated and reported Title I evaluation information. Some states currently gather data at the individual student level from the LEA's and then employ central processing; other states rely entirely on their LEA's to analyze their own data. Some states have automated virtually the entire evaluation process, using scoring services and automatic data processing; others rely on information that is scored and analyzed entirely by hand. In states that have both very large and very small LEA's, differences in procedures vary tremendously even within a state.

There are many reasons for this wide variety of practice across the national Title I system. Capabilities and support systems vary widely. Different philosophies and priorities are used to set policies. The number of Title I students in a district may range from the tens to the tens of thousands, and the amounts of Title I grants might range from the tens of thousands to the tens of millions of dollars. Thus, it is difficult to envision a unique correct way to do things; the system must be designed to function efficiently with alternative methods for data handling, processing, analyzing, and aggregating.

Although the evaluation methods and capabilities of the states and localities that have implemented the RMC Title I evaluation models vary widely, it appears that the flow of achievement data through the evaluation systems involves certain tasks which are common to all states and localities. These tasks are (1) the selection and administration of tests, (2) the scoring of instruments and conversion of scores, (3) the analysis of data, and (4) the aggregation of data.

Selection/administration of tests: At this initial phase of an evaluation, several steps are important: (1) the proper test must be selected; (2) the correct level of it must be administered; (3) the test administration procedures must be standard; and (4) the testing conditions must be appropriate.

The "proper" test is, foremost, one which measures what is being taught. Much research (Armbruster, Stevens, and Rosenshine, 1977; Bianchini, 1976; Hoepfner, 1976; Stearnes, 1977; and Tallmadge, 1977) has highlighted the

degree to which different standardized tests emphasize different subskills within a skill area. Title I evaluators are advised that the more closely their tests correspond to the project objectives, the more relevant the scores will be for detecting student growth in the project (Fagan and Horst, 1978). This fairly common-sense notion is often disregarded as other factors influence test selection.

The "proper" test has empirical normative data on children similar to the children in the project and gathered at dates in the school year corresponding to the pre and post-test dates for an evaluation using model A. Also, the norm data are from a representative national or local sample of children.

Using the correct level of the test means administering one on which the fewest children possible score either at the "chance" level or at the top score. This is important because a preponderance of the former, placing many students at the "floor" of the test, artificially inflates the group's pretest average, thereby overstating their status before the project. An estimate of their gain due to the project would then be underestimated. Similarly, if the students "top out" on the post-test, the group's status after the project is underestimated, and the resulting gain figure is again too small. Of course, the mismatch of test level with student skill levels can also affect evaluation results in other ways; the important consideration is that students' performance levels be reflected as accurately as possible. Use of the wrong level of the test precludes this (Roberts, A.O.H. 1978; Roberts, S., 1978).

All tests have standardized procedures outlined for their use. Even "homemade" instruments have instructions for administering them. Such procedures may include timing, use of practice items, degree of assistance from the test proctors, etc. Furthermore, the testing conditions must be good. For example, the rooms should be quiet; the settings and times for testing project and "comparison" group children should be similar. (Horst, 1978; Tallmadge and Roberts, 1978). In order for students' test scores to be comparable to those of others, especially to the norm data, the outlined procedures must be followed (Horst, 1978).

Scoring of instruments/Conversion of scores This step involves determining the test score for each student. Occasionally, it is done by hand but, more often, a scoring service is used. Student answer sheets are sent away to a firm which will return lists of students tested, the raw score for each, and requested score conversions. However, problems may arise with lost answer sheets, incorrectly coded ones, damaged pages, or errors in the conversion of scores.

Due to the mathematics required by the analytical procedures in an evaluation, students' raw scores must be converted to a standard score metric. In most cases, the preferable standard score metric for the computations is one which incorporates characteristics of the national distribution of scores for the age group--the normal curve equivalent (NCE). In order to derive that figure, as many as three or four separate conversions may be necessary for each student's score. A typical sequence would be to convert the child's raw score to the publisher's standard score, then to a national rank or percentile, and finally to an NCE.

Some scoring services can provide all of these scores; all can provide the percentile equivalents. Typically, there is a charge for each additional score requested, so the use of a service for all of the conversions would be expensive for large projects. Scores from scoring services are usually more accurate than scores that have been manually tallied and converted, however.

Data analysis. This is the phase in the Title I evaluation where data from individuals are combined into project-level statistics. NCE gains, describing the effectiveness of the Title I project in contributing to the students' learning above and beyond what is expected from the "regular" curriculum, are calculated. Each model prescribes the appropriate analytical techniques, which range from fairly straight-forward computations in the case of Model A to complex statistical manipulations in Models B and C.

Data aggregation. This is the final step in the Title I evaluation system prior to the actual reporting of the evaluation results. Errors made at this stage of the evaluation are not likely to be serious, since the data on which

the aggregations are based, the project-level NCE gains, are usually accessible for later double-checking. This phase usually takes place at the SEA level.

PROBLEMS ENCOUNTERED IN IMPLEMENTING THE MODELS

In using the new models, SEA's and LEA's had problems and made mistakes in three areas: procedural, referring to adherence to suggested rules of the models; clerical, referring to recording, translating, and calculating; and analytical, referring to technical and statistical problems.

Procedural. The first procedural problem is in the selection of Title I students. The requirement of Model A that selection of Title I participants be based on data other than from the pretest is the requirement that has been most often ignored. Even among the states which were the first to attempt the models, states with extremely competent Title I personnel, the proportion of LEA's still selecting on the pretest exceeded half of those reporting.

SEA personnel were beset with questions from the LEA's on how to follow the model's rules without deviating from present practices or perceived requirements. Various LEA and SEA personnel advised their LEA's of methods in which to choose students without invalidating the evaluations, but the LEA's seldom tried the advised methods or applied the methods validly. For example, one evaluator suggested choosing students based on a combination of standardized achievement test scores, absenteeism, a teacher's estimate of the student's achievement, an estimate of underachievement, and an estimate of motivation and health. When the reports were returned, it was discovered that most school districts had used the pretest standardized test score combined with an estimate of underachievement which had been obtained by the subtraction of the student's pretest score from the class average. Thus, only a single criterion was used and the choice of individuals to be placed in the Title I group was based solely on the pretest.

The second most common procedural problem lies in the administration of norm-referenced tests at the proper time in Model A. Comparisons between the Title I group and the test publisher's norms are most valid when based on real data points, so the tests should be given during a four or six-weeks period spanning the test publisher's main norming date. This was frequently not done by the LEA's that we studied. LEA personnel generally wish to test

as early in the school year as possible (e.g., in the middle of September). However, the midpoint of a publisher's test norms may not occur until November.

Another procedural problem is that of communicating to local administrators the meaning of the new NCE metric. According to a number of state leaders, parents accept the idea with little hesitation, but administrators, especially superintendents, resist strongly. The use of a variety of metrics for sharing results with interested parties will help alleviate this problem.

A final procedural matter is the fact that administrators consider local funding allocations for testing when they choose Model A, B, or C. The requirements of Model C include the testing of comparison students from among those not in the project. Therefore, if the district already has budgeted funds for testing once a year, and selection of Title I students on that test is acceptable, then Model C is a logical choice. But if the money for testing has not been allocated, the choice can just as logically be Model A.

Clerical. The translation of a raw score into any other score is fraught with error. In one state, data from more than 93% of the LEA's contained at least one table look-up error. Another state director reported that the majority of errors in his workshop exercises stemmed from the inappropriate use of the same norm table for both pretest and posttest score conversions in spite of very obvious table titles. Where yet another had modified the format of the publishers' norm tables, the error rates were considerably reduced--but still excessive.

Another problem is that often the gain score for an individual turns out negative, and negative numbers appear to be an anathema to proper calculations. One evaluator was so upset by negative scores that he ignored every one in averaging gains in his project.

The assignment of a pretest or posttest score to every individual is only part of the process--the two tests must be matched for each individual. In large districts the matching is generally done on a computer. Here several difficulties appear: a matching program must be written; bad coding or punching takes a toll of properly matched individuals; a single unmatched card in a sorted file makes the entire remaining ones mismatched and makes

the evaluation worthless if left undiscovered; students change the spelling of their names; and inconsistencies appear in using the last name first.

In smaller districts, the match is carried out by hand. Here the human intelligence can remedy many of the problems: one can see that "Bill Thorne" on the pretest is "Thorne, W." on the post, for example. A knowledgeable individual might remember that a school boundary was recently changed, and half the students are in the neighboring school, or that the Jones children oscillate regularly between two schools. Thus the increased difficulty in matching by hand is offset by the increased frequency of matches.

A comparatively minor error in reporting is the failure to include the day of the month on which testing was accomplished. Under Model A, a user of the ITBS would be expected to test within a two-week interval on either side of April 28th. If he merely reported that the test was administered in April, the state evaluator cannot assume that the test was given at the proper time; it could have been on the first of April. The error is minor in that it can be corrected easily with a change in the reporting requirements.

Analytical. A variety of technical questions dealing with the statistical and psychometric aspects of the system continue to plague the evaluators. First, are evaluators jeopardizing the accuracy of evaluation results by testing once a year in the spring? If students forget a great deal during the summer, perhaps they would show greater gains if they were tested in both fall and spring. Second, when students repeat a grade, what pretest score should be used--the first pretest or the second? What norms should be used for such students?

Some state personnel note that the correction of an LEA's evaluation error may result in a lowered gain estimate. They recognize, however, that no one wants to be fooled into assuming successes in remediating children's educational problems if the remediation has not occurred.

The sample of individual project reports perused in the states were rated on their quality of evaluation. Those projects with the "best" evaluation showed low but positive gain scores for the Title I group. (The correlation between evaluation quality and size of gain was -.25.) Those

projects which appeared to be evaluated correctly showed a modest positive impact from the program.

SUGGESTIONS FOR IMPROVEMENTS

When the directors and evaluators in the SEA's and LEA's identified the problems outlined above, they also suggested a variety of solutions. Some approaches had already been tried out, and other possibilities arose in discussions.

People we have worked with and other interested parties have suggested ways USOE can help SEA's and LEA's follow more completely the procedures outlined in the evaluation and reporting system. Their suggestions are summarized below according to the same categories used in preceding sections: procedural, clerical, analytical.

Procedural. Most of these apply to the general implementation rules and other administrative areas.

1. Provide a detailed special handbook on the implementation of each model. The handbook should be very elementary, in step-by-step flow chart fashion, with plenty of concrete examples of documents and approaches which have worked.

2. Emphasize reduced testing requirements with the proposed models. Too many school personnel are too worried about too many tests. Specify, for each model, the minimum testing possible.

3. To encourage the proper administration of tests, encourage those districts using once-a-year spring testing to have the teacher from the next higher grade give the test. To the third grade teacher testing the second grade students at the end of the year, accuracy would be paramount since next year he would have those very students and would, supposedly, welcome accurate test scores in their folders.

4. Give more guidance regarding test selection. Many studies have demonstrated the importance of test content for detecting student growth in specific skill areas. Certain tests may be more sensitive than others to the skills content of many Title I programs.

5. Add more information to the handbook on out-of-level testing. Most individuals still feel very uncomfortable attempting to implement functional-

level testing although they recognize the necessity with some students. Prepare a detailed checklist which states can give to LEA's showing the effort involved in ordering, the logistics of testing, and the scoring and translating of scores. Discuss the pay-offs for these extra steps in terms of an increase in accuracy of the scores. Point out and discuss common misconceptions, such as the beliefs that functional-level testing will result in the pretest average being lower, that it will result in choosing the wrong students, that it will give inaccurate estimates of gain, or that it will compare students at one grade unfairly with those at the next lower grade.

6. Provide guidelines on what action to take if the raw score from an out-of-level test administration leads to a converted score too low to be included in the percentile conversion table.

7. Communicate the results of the data collected in this study to publishers, especially the information about the needs for norming earlier in the year, and for less confusing norms tables.

8. Remove the suggestion in current documentation that two-thirds of the project should take place between the pre and post-tests. The incidence of failure to follow this requirement is negligible, and should diminish to zero as districts move to appropriate testing dates. The requirement leads to reporting of non-informative data and to problems when schools have a provision for students' return to their regular classrooms after they have mastered a certain body of material.

9. Investigate the conditions under which combining results from different grades is appropriate. If a comparison group for Model C is too small within a particular grade, some addition of non-Title I students from the next higher grade might be possible although technically they are not in the project. In Model A, adding the two Title I students in 8th grade to those in 7th might reduce the trauma attendant to finding an average of a 10 NCE loss (due to small size and unstable data).

10. In the light of the high rate of errors in table reading, consider the use of a raw score reporting system. In the absence of mechanical aids to table look-up, provide simpler score conversion tables to users and

communicate the problem to test publishers. Some publishers have already provided vastly improved tables for those states in which the evaluator has insisted upon them.

11. Provide assistance, both in guidelines and in a computer program, for the plotting of pretest against posttest for all models. The visual inspection of the scatterplots could be expected to indicate floor and ceiling effects, gross non-linearity of relationships, and the presence of unexpectedly high or low "gainers".

12. Classify evaluation errors by severity so that when two implementation suggestions conflict, the Title I co-ordinator can have some guidance in making a choice between the two.

13. At present, the federal government publishes excellent technical reports on the RMC models, but SEA's and LEA's often do not obtain the reports because they are expensive or because the SEA's and LEA's do not know that the reports exist.

Make these reports available to SEA's at no cost. The SEA's could then distribute them to LEA's.

14. Show some examples of how program objectives could be stated under the new reporting format.

For example, one state requests that the LEA Title I director estimate the NCE gain to be achieved for each grade within each project. If the third grade reading project at Memorial Elementary has a reputation as the best, then a seven NCE gain may be the objective. If it is the worst, then one-half an NCE might be appropriate.

15. Prepare examples to help SEA's and LEA's communicate evaluation results, (including use of NCE's and percentiles) to Title I parents, school boards, and teachers and administrators.

16. Consider alternative methods of tying a cost figure to a project. Since districts generally spend about 75% of their budget for instructional personnel direct costs, the reporting of only those costs may increase the accuracy and decrease the reporting burden of the cost estimate.

Clerical. Clerical suggestions refer to recording, translating, and calculating processes.

1. Instruct LEA evaluators, when looking up average standardized scores in publishers' norm tables, to use the individual percentile norms table, not the school norms table.

2. Remove any requirements for data point interpolation at the district level. Interpolation appears to be more error-prone than it is worth.

3. Send copies of an exemplary testing-dates chart to all interested parties. For example, the forms used in one state include a chart which assists district personnel in avoiding test administration date errors.

4. Add a requirement for the project report to provide the average selection NCE, where possible, and the average pretest NCE. This will allow an easy edit check to see if the selection was based completely on the pre-test, and if the most needy students were chosen.

5. Revise the percentile-to-NCE conversion tables so that they will be easier to use. Perhaps they could be placed in groups of ten, with clear lines to demarcate the columns.

6. Develop optical scanning forms and software for the analysis of scores from major tests. If a state, or large LEA, decides to centralize the scoring process rather than have the local classroom teachers score their own papers, this is the solution with the greatest long-range potential, though perhaps the highest initial expense. Furthermore, although the accuracy and speed of data processing will be vastly improved, the system must also have a set of appropriate error-detection methods built in.

7. Encourage LEA's to score some tests by hand, even if they have engaged a scoring service, in order to check the accuracy of the service.

8. Provide guidance regarding the use of automated data processing as often as possible for the various score conversions and manipulations necessary in the system. Encourage other approaches, too, to preserve the integrity of the data:

(a) Staff should try to perform score conversion activities in teams, with people double-checking the work of others whenever possible;

(b) Raw data should be stored (or sent to the LEA or SEA) to enable

later checking of a data sample for the correctness of the tables used, score conversions, etc.;

(c) Tables should be re-formatted at the local level to allow for easier reading (e.g., 2-column tables are much easier to work with than multi-column tables); and,

(d) Staff responsible for performing the score conversions should be trained in the use and interpretation of standardized tests, so that they will understand the meaning of the scores and will be able to detect obviously wrong, inappropriate, or out-of-range scores.

A pre-programmed calculator or computer can also facilitate the data analysis. For small LEA's and SEA's without access to larger computer facilities, a set of programs and implementation materials is being developed for use with hand-held programmable calculators available from $89 and up. These programs can certainly aid the evaluator but may not offer a complete set of diagnostic features (e.g., a plot of pretest against posttest scores) of the types which have been suggested in this paper. Properly designed microprocessor-based software appears to offer more acceptable and cost-effective provisions for the extensive amount of data checking and editing which should be done by the evaluator.

Analytical. The suggestions in this section refer to technical characteristics of the models.

1. Clarify the severity of the regression hazard when two-stage selection takes place. For instance, a potential Title I treatment group of 100 students may be identified by teacher referral. Then a pretest is given to those 100, and the 95 students most in need are given treatment. The regression effect is considerably less than if the five students most in need were chosen.

2. Investigate further the trade-off in comparing project effects from spring to spring versus fall to spring. Once-a-year testing is much easier on the budget and school time, but this may be offset by the loss of students between school years and the students' loss of knowledge over the summer.

3. Assist evaluators in the objective identification of "outliers" (student data so extreme that they are likely in error). Outliers can

significantly affect estimates of project effect especially when Model C is used. For instance, in plotting a Model C implementation, one district evaluator was surprised to find two comparison group individuals who were at the top of the distribution on the pretest and almost at the bottom on the posttest. Examination of their scores revealed that they had scored at the 99th percentile on the reading comprehension pretest, without missing a single item. In contrast, their vocabulary scores were at the 1st percentile. Clearly some error had been made (by the scoring service, perhaps), and the individuals' scores were dropped from the analysis. The result was that the estimate of project effects changed from negative to positive.

CONCLUSIONS

It is apparent that the process to change LEA evaluation activities to conform with those prescribed by the evaluation models is laborious and, by necessity, iterative. States we visited had staff pursuing this goal for as long as two years, and many reported that more work is still needed.

Probably the most pervasive administrative problem for SEA's is that they do not know and therefore cannot review what actually happens in the LEA's. Of course, this is much more than just an evaluation problem, but it greatly affects evaluation data. A test may have been given on other than the reported dates (in one of the states, the test was supposedly administered on a Sunday); the test administrator may have ignored the directions for proper test giving; tests designed for group administration may have been given individually or with relaxed time limits; tests may have been administered to the wrong individuals.

Reliance on information that is scored, transformed and analyzed manually appears to be a major threat to the validity of reported data. Though conversion to automatic data processing wherever possible in the evaluation system does not promise to be a total panacea, it seems a promising first step.

Although the problems in instituting the RMC models are legion, both state and local district personnel have found the models to be extremely useful. The models allow state personnel to compare data across districts. Because of

this, some states are already using the evaluation data to identify especially effective programs in specific LEA's and most states plan to start using the data in this way. Other states have already sponsored or encouraged investigations to determine which Title I program features appear to be positively correlated with achievement gains.

Another advantage to these models is the existence of standards which states can use for advising and monitoring their LEA's evaluation activities. Some SEA's believe that the lack of such information historically left them with little basis for insisting upon specific LEA evaluation practices.

State personnel have also noted the benefits of greater attention to achievement tests--their content, use, selection, etc. For example, the fact that the models recommend specific procedures appropriate to the test being used has prompted evaluators to look more deeply into the characteristics of their tests.

Local personnel welcome the possibility of comparing the outcomes of their efforts to those of districts they know to be similar. In addition, LEA's have indicated that the explicit recommended procedures of the new models are less burdensome than the former federal mandates, which were unclear.

In summary, we believe that the problems encountered in instituting the RMC models can be solved. We submit that the advantages of using the models make the efforts on the part of LEA's and SEA's to accomodate themselves to the necessary changes well worth-while.

Armbruster, B.B., Stevens, R.J. and Rosenshine, B. Analyzing Content Coverage and Emphasis: A Study of Three Curricula and Two Tests. Urbana, Illinois: University of Illinois, 1977.

Bianchini, J.C. Achievement Tests and Differentiated Norms. Berkeley, California: Educational Testing Services. Paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, Virginia, 1976.

Fagan, B.M. and Horst, D.P. Selecting a Norm-Referenced Test. Mountain View, California: RMC Research Corporation, 1978*.

Hoepfner, R. Achievement Test Selection for Program Evaluation. Santa Monica, California: System Development Corporation. Paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, Virginia, 1976.

Horst, D.P. and Wood, C.T. Collecting Achievement Test Data. Mountain View, California: RMC Research Corporation, 1978*.

Roberts, A.O.H. Out-of-Level Testing. Mountain View, California: RMC Research Corporation, 1978*.

Roberts, S.J. Test Floor and Ceiling Effects. Mountain View, California: RMC Research Corporation, 1978*.

Stearns, M.S. Evaluation of the Field Test of Project Information Packages, Volume I Summary Report. Menlo Park, California: SRI International, 1977.

Tallmadge, G.K. Title I Evaluations: Comparable Outcome Measures for Dissimilar Instructional Treatments? Mountain View, California: RMC Research Corporation. Paper presented at the 27th Annual Conference of Directors of State Testing Programs, Princeton, New Jersey, October 1977.

Tallmadge, G.K. and Roberts, A.O.H. Factors That Influence Test Results. Mountain View, California: RMC Research Corporation, 1978*.

---

*Produced under USOE contract 300-76-0316. Copies are available from the U.S. Office of Education, 400 Maryland Avenue, S.W., Room 3049, Washington, D.C., 20202.